Nyelvadaptáció a többszavas kifejezések automatikus azonosításában

Nagy T. István[1], Vincze Veronika[2]

[1]Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
Szeged Árpád tér 2., e-mail: nistvan@inf.u-szeged.hu
[2]MTA-SZTE, Mesterséges Intelligencia Kutatócsoport,
Szeged, Tisza Lajos körút 103., e-mail: vinczev@inf.u-szeged.hu

Kivonat Ebben a munkában bemutatjuk gépi tanuláson alapuló nyelvfüggetlen rendszerünket különböző nyelvű félig kompozicionális szerkezeteknek automatikus azonosítására. A módszer szintaktikai elemzésen alapuló jelöltkiválasztó megközelítést használ, mely a lehetséges félig kompozicionális szerkezetekről egy gazdag jellemzőtérre támaszkodó gépi tanuló megközelítés segítségével hoz döntést. Az eredményekből kiderül, hogy a más nyelvekből származó adatok is pozitív hatással bírnak a szerkezetek azonosítására.
Kulcsszavak: többszavas kifejezések, többnyelvűség, nyelvadaptáció, 4FX

1. Bevezetés

Ebben a munkában bemutatjuk gépi tanuláson alapuló nyelvfüggetlen rendszerünket a többszavas kifejezések egy osztályának, a félig kompozicionális főnév–ige szerkezeteknek (FX-eknek) azonosítására. Megközelítésünk alapjául egy korábban ismertetett, angol és magyar nyelvre kifejlesztett kétlépéses módszerünk szolgál [1]. Első lépésben a lehetséges FX-jelölteket nyerjük ki a szövegből szintaktikai szabályok segítségével, majd a továbbiakban döntési fákon alapuló gépi tanuló módszerekkel határozzuk meg, hogy a jelölt ténylegesen többszavas kifejezés-e. Módszerünket a 4FX nevű párhuzamos korpuszon [2] teszteljük: a korpusz angol, spanyol, német és magyar nyelven tartalmaz szövegeket, melyekben kézzel be vannak jelölve a félig kompozicionális szerkezetek. A korpuszra támaszkodva bemutatjuk nyelvadaptációs kísérleteinket is, melyek igazolják, hogy az eltérő nyelvekből származó adatok is pozitívan hatnak a rendszer eredményességére.

2. Kapcsolódó munkák

Számos különböző módszer született már az FX-ek automatikus azonosítására különböző nyelveken. Az angol nyelvű kutatások [3, 4] során jellemzően csupán az adott szövegben előforduló ige–tárgy párokra fókuszálnak az FX-ek detektálása közben. Ugyanakkor a nem angol nyelvű kutatások során, mint például

holland [5], alapvetően ige-prepozíció-főnév szerkezeteket vizsgálták. Az alapvetően csak statisztikai jellemzőkre építő megközelítések [4,5] hatékonysága erősen korlátozott, hiszen ezen módszerek a ritkán előforduló FX-eket igen nehezen tudják detektálni. Ugyanakkor ezen szerkezetek nagy többsége meglehetősen ritkán fordul elő egy adott korpuszon belül [6].

A szabályalapú rendszerek [7, 8] jellemzően sekély nyelvi információkat felhasználva azonosították az FX-eket automatikusan, míg Vincze és társai [9] a SzegedParallelFX párhuzamos korpuszon mutatták be szabályalapú módszerüket magyar és angol nyelvű FX-ek azonosítására.

Statisztikai, valamint nyelvi információkat egyaránt felhasználó gépi tanuló megközelítéseket [3, 10, 11] szintén alkalmaztak FX-ek detektálására. Mindegyik módszer alapvetően ige + főnév párokat osztályoz aszerint, hogy az adott szerkezet FX-e vagy sem. Ugyanakkor a Nagy T. és társai [11] által bemutatott módszer már nem csak a szövegben előforduló ige-tárgy párokra fókuszált, hanem különböző szintaktikai jellemzők alapján automatikusan kinyert főnév + ige párokat osztályoz gazdag jellemzőtérre támaszkodó gépi tanuló módszerük alapján.

Az általunk bemutatott FX-azonosító megközelítés a [11] által demonstrált módszeren alapszik, melyet oly módon módosítottunk, hogy az képes legyen FX-ek automatikus azonosítására különböző nyelveken. Ismereteink szerint egyetlen korábban ismertett módszer sem tett kísérletet egy általános modell létrehozására, mely képes különböző nyelvű FX-ket automatikusan azonosítani.

3. Kísérletek

Célunk minden egyes FX azonosítása a 4FX korpusz [2] minden egyes nyelvén, gépi tanulási módszereket alkalmazva. Ehhez egy angol és magyar nyelvre már megalapozott módszerből [1] indulunk ki, melyet németre és spanyolra is átalakítunk az adott nyelv sajátságainak megfelelően.

Kísérleteink során azt is megvizsgáljuk, hogy a különféle nyelvek hogyan hatnak egymásra a tanítás során, így szükségesnek bizonyult egy nyelvfüggetlen reprezentáció kialakítása. Ezáltal a doménadaptációs eljárásokhoz hasonló nyelvadaptációs technikákat is ki tudunk próbálni a korpuszon.

3.1. A jelöltek kiválasztása

Az FX-jelöltek kiválasztásában [11] módszerét követjük, azaz a szintaktikailag elemzett szövegből kinyerjük az előre megadott függőségi kapcsolatok egyikét alkotó szavakat, majd ezt követően bináris osztályozás segítségével eldöntjük, hogy azok ténylegesen FX-ek-e.

Az angol, német és spanyol szövegek függőségi elemzéséhez a Bohnet parsert használtuk [12], az angol esetében a CoNLL-2008 korpuszon [13], a német esetében a TIGER treebanken [14], míg a spanyol esetében az IULA treebanken [15] tanítva. A magyar szövegek elemzéséhez a magyarlanc 2.0-t [16] alkalmaztuk a Szeged Dependencia Treebanken [17] tanítva.

Minthogy a különböző nyelvű treebankek eltérő függőségi címkéket használtak, így a 4FX korpusz különböző nyelvű változataiban is eltérő címkék szerepelnek ugyanannak a nyelvtani viszonynak a jelölésére, például az ige-tárgy kapcsolat jelölése az angolban a `dobj`, a németben az `OA`, a spanyolban a `DO` és a magyarban az `OBJ` címkével valósul meg. Így egységesítettük a nyelvtani viszonyok jelölését a nyelvek között, hasonlóan az univerzális dependenciaannotációhoz [18], azonban mi csupán az FX-ekre vonatkozó viszonyokkal foglalkoztunk.

A jelöltkinyerő fázisban FX-jelöltnek tekintettünk minden egyes ige–tárgy, ige–(passzív) alany, ige–adpozíciós frázis és főnév–igenévi módosító szókapcsolatokat.

3.2. Egységesített jellemzők

Gépi tanulási kísérleteinkhez [1] angolra és magyarra kifejlesztett módszereit vettük át, és adaptáltuk németre és spanyolra, bevezetve ezáltal néhány nyelvspecifikus jellemzőt. A nyelvfüggetlen jellemzők mellett, melyek az FX-ek nyelveken átívelő sajátságait tükrözik, az egyes nyelvekre saját jellemzőket is megadtunk, mivel az eltérő nyelvek eltérő nyelvtani sajátságokkal bírnak: például a főnevek nyelvtani neme jellemzőként szerepel a spanyol és a német nyelv esetében, ugyanakkor az angol és magyar esetében erre a jellemzőre nincs szükség.

Az egyes nyelvekre használt jellemzőket rendre megfeleltettük egymásnak, lehetővé téve ezzel a nyelvadaptációt. Például a leggyakoribb igei komponensek fordításait minden nyelvben megfeleltettük egymásnak, rendezett négyeseket képezve: *take – nehmen – tomar – vesz*. A lexikai jellemzőkhöz hasonlóan, a szintaktikai és morfológiai jegyeket is egységes nyelvfüggetlen alakra hoztuk.

Morfológiai jellemzők: megvizsgáltuk a főnevek szótövét, és bináris jellemzőként felvettük, hogy a főnév igéből képzett-e. Az FX-jelöltek tagjainak szófaját is felvettük mint jellemzőt, amennyiben az egyezett egy előre megadott lehetséges szófaji mintával, mint például `ige + főnév`.

Néhány nyelvfüggő jellemzőt is megadtunk minden egyes nyelv esetében. Az angol nyelvű morfológiai elemzés megkülönbözteti a főigéket és segédigéket, így tehát a *do* és *have* igék esetében azt is szerepeltettük, hogy azok főigei vagy segédigei használatban fordulnak elő az adott mondatban, mivel mindkét ige gyakran fordul elő FX-igeként is. A magyar nyelv morfológiailag gazdag lévén számos morfológiai jellemzőt vettünk fel a szavak morfológiai elemzése alapján mint például az igék módja, a főnevek esete, a birtokos száma és személye és a birtok száma. A nyelvtörténetileg igéből származtatott főneveket, melyeket a morfológiai elemző nem kezelt képzésként, szintén külön jelöltük.

A német és spanyol nyelv esetében a főnevek nyelvtani nemét is felvettük jellemzőként, mivel képzőiknek köszönhetően az FX-eket alkotó főnevek gyakran nőneműek ezekben a nyelvekben. Ezen túl, a német nyelvre felvettünk egy újabb jellemzőt, mely azt jelöli, hogy a főnév összetett szó-e vagy sem. A spanyol melléknévi igeneveket külön is megjelöltük végződésük alapján, mivel a morfológiai elemzés nem különbözteti meg a mellékneveket és a melléknévi igeneveket, azonban míg a melléknévi igenevek szerepelhetnek FX-ek részeként, addig a melléknevek nem.

Felszíni jellemzők: Mivel az FX-ek főnévi komponenseit gyakran képzik igéből, így a tipikus igeképzőket bi- és trigramként kezelve megvizsgáltuk, hogy az FX-jelölt főnévi komponense az adott bi- vagy trigramban végződik-e. Az FX-jelöltek szószámát szintén felvettük jellemzőként.

Statisztikai jellemzők: A jelöltkinyerő módszerrel kigyűjtöttük 10 000 angol Wikipedia-oldalból a lehetséges FX-eket, majd feljegyeztük ezek előfordulási gyakoriságait. Amennyiben az FX-jelölt megegyezett az egyik, listában szereplő egységgel, akkor jellemzőként felvettük a gyakoriságát is.

Lexikai jellemzők: Mivel általában a leggyakoribb igék fordulnak elő FX-igeként, ezért minden nyelvben kiválasztottunk 15 gyakori igét, és megvizsgáltuk, hogy az FX-jelölt igéje megegyezik-e velük. A nyelvközi méréseinkhz egyesítettük az egyes nyelvek FX-listáit, és minden egyes igét lefordítottunk mind a négy nyelvre, függetlenül attól, hogy az adott nyelvű ige éppen benne volt-e a leggyakoribb 15-ben. Így igenégyeseket kaptunk, mint például *make - machen - fazer - tesz*. Az így létrehozott lista összesen 29 igenégyest tartalmaz.

A főnevek lemmáját is jellemzőként hasznosítottuk. A parserek tanításához használt treebankekből gyűjtöttük össze az FX-ekben található főneveket.

A fentieken kívül lemmatizált FX-listákat is hasznosítottunk jellemzőként. Az angol és magyar esetében a SzegedParalellFX korpusz [19] megfelelő részéből kigyűjtött FX-eket használtuk, míg a német esetében a német PP-igei kollokációkat tartalmazó listából [20] szűrtük ki az FX-eket. A spanyol esetében az ige-főnév párokat lexikai függvények alapján kategorizáló szótár anyagából [21] indultunk ki.

Szintaktikai jellemzők: Jelöltkinyerő módszereink elsődlegesen a főnév és az ige közti szintaktikai kapcsolatra építenek, azonban a szintaktikai kapcsolatok a tényleges FX-ek kiválasztásában is hasznosíthatók. Így tehát a 3.1. részben bemutatott függőségi viszonyokat használtuk fel szintaktikai jellemzőként. Amennyiben a főnév rendelkezett névelővel, azt is jelöltük a jellemzők között.

A 1. táblázat mutatja, mely nyelvekre mely jellemzőket alkalmaztuk.

3.3. Gépi tanuláson alapuló osztályozás

[11] már korábban bemutatta, hogy ezen a feladaton a döntési fákon alapuló megközelítések teljesítenek a legjobban, ezért a WEKA gépi tanuló csomagban [22] található J48 döntési fa algoritmust tanítottuk a fentebb leírt jellemzőkészleten. Modelljeinket tízszeres keresztvalidációval értékeltük ki a korpusz minden részén.

Mivel a szintaktikai elemzésen alapuló jelöltkiválasztó megközelítés nem képes az összes manuálisan annotált FX-t kinyerni, ezért a kimaradt FX-eket téves negatívként kezeltük a kiértékelés során.

Mind a négy nyelven esetében egy kontextusfüggetlen szótárillesztési megközelítést is alkalmaztunk alapmódszernek, ahol a 3.2 fejezetben ismertetett FX-listákat alkalmaztuk. A szótárban található FX-eket abban az esetben jelöltük az adott szövegben, amennyiben azokat a szintaxisalapú jelöltkiválasztó megközelítés előzetesen kinyerte és a szövegben előfordultak.

1. táblázat. Nyelvfüggetlen és nyelvfüggő jellemzők

Jellemző	**Nyelvfüggetlen**	**Angol**	**Német**	**Spanyol**	**Magyar**
Felszíni	•	•	•	•	•
Szintaktikai	•	•	•	•	•
FX-listák	•	•	•	•	•
Igelisták	•	•	•	•	•
Főnévlisták	•	•	•	•	•
Szófaji minta	•	•	•	•	•
Igei szótő	•	•	•	•	•
Főnévképző	•	•	•	•	•
Statisztikai	–	•	–	–	–
Segédige	–	•	–	–	–
Összetett főnév	–	–	•	–	–
Nem	–	–	•	•	–
Melléknévi igenév	–	–	–	•	–
Agglutináló morfológia	–	–	–	–	•
Nyelvtörténeti képző	–	–	–	–	•

3.4. Nyelvadaptáció

Nyelvadaptációs vizsgálatainkban a doménadaptációhoz hasonló módszert használtunk. A doménadaptációs technikák alkalmazása leginkább akkor sikeres, ha egy adott doménből viszonylag kevés adat áll rendelkezésre, azonban egy másik doménből sok adathoz férünk hozzá. Esetünkben a különböző nyelveket tekintettük különböző doméneknek, így megvizsgálhattuk, hogy az eltérő nyelvekből származó adatok hogyan befolyásolják az FX-ek azonosításának eredményességét.

Többféle mérést is elvégeztünk a rendelkezésre álló korpuszon. Először mind a négy nyelven tízszeres keresztvalidációval tanítottuk és értékeltük ki a rendszert. Ezután minden egyes nyelvpár esetében keresztméréseket is alkalmaztunk, azaz a forrásnyelvet használtuk tanító adatbázisként, és a célnyelven értékeltük ki a rendszer teljesítményét. Végül nyelvadaptációs méréseket is végrehajtottunk minden egyes nyelvpár esetében, ahol a tanító adatbázis a forrásnyelvi adatok mellett a célnyelvből származó adatokat is tartalmazott kis mennyiségben, a kiértékelés pedig a többi célnyelvi adaton valósult meg. Összehasonlítási alapként szótárillesztéses méréseket is végeztünk minden egyes nyelvre.

A keresztmérésekhez elvégzéséhez az FX-jelöltek egységes reprezentációja szükséges. Ugyanakkor, ahogy a 1. táblázat is mutatja, az FX-ek különböző nyelveken való automatikus detektálásához nyelvspecifikus jellemzőket is definiáltunk, ezért az alap jellemzőkészletet kiegészítettük az összes nyelvspecifikus jellemzővel.

A nyelvadaptáció során egy egyszerű megközelítést alkalmaztunk (ADAPT): tízszeres keresztvalidációt alkalmaztunk, ahol a célnyelvből 10%-ot használtunk tesztelésre, míg a maradék 90%-t a tanítás során hozzáadtuk a forrásnyelv tanító

halmazához. Forrásnyelvnek a nyelvek összes lehetséges kombinációját alkalmaztuk, ami nem tartalmazta a célnyelvet.

A nyelvadaptáció kiértékelése során a gépi tanuló megközelítésünket a forrásnyelv és a célnyelv tesztelésre fel nem használt részének unióján tanítottuk, a kapott modelt pedig tízszeres keresztvalidációval értékelük ki a célnyelven. A keresztvalidáció során minden alkalommal a célnyelv 10%-át használtuk tesztelésre, míg a maradékot tanításra.

Az angol, német, spanyol és magyar nyelvekre végzett nyelvadaptáció eredményei a 2., 3., 4., illetve 5. táblázatokban találhatók.

4. Eredmények

Az alkalmazott módszerünk az indomén mérések során relatíve azonos eredményt ért el a korpusz magyar és angol részein 65 körüli F-mértékkel, míg spanyol és magyar nyelveken 50-es F-mértéket meghaladó eredményt ért el. A nyelvadaptáció eredményei minden korpuszon meghaladták a szótárillesztését, sőt, a spanyol nyelvet kivéve, a keresztmérések is hatékonyabbnak bizonyultak a szótárillesztésnél.

A korpusz angol részén elért eredményeket a 2. táblázat mutatja, ahol a nyelvadaptáció során mind a három másik nyelv képes volt javítani az eredményeken. A nyelvadaptáció 66,30-as F-mértéket elérve, abban az esetben bizonyult a leghatékonyabbnak, amikor a 4FX korpusz spanyol és német részének unióján tanítottuk, míg a keresztmérések esetében a német korpuszon tanított modell bizonyult a leghatékonyabbnak. A keresztmérések átlagos eredményei 7,99 F-mértékkel bizonyultak jobbnak a szótárillesztésnél, míg a nyelvadaptáció eredményei jelentősen meghaladták azt.

2. táblázat. Kísérleti eredmények a angol részkorpuszon. EN: Angol. DE: Német. ES: Spanyol. HU: Magyar. ADAPT: nyelvadaptáció. CROSS: keresztmérések.

Nyelvek				**ADAPT**			**CROSS**		
EN	**DE**	**ES**	**HU**	**Pontosság**	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Szótárillesztés				83,85	19,71	31,92	83,85	19,71	31,92
♠				**78,81**	**55,82**	**65,35**	–	–	–
♠			•	77,93	**56,60**	65,58	75,65	27,36	40,18
♠		•		79,18	56,13	65,69	67,84	21,23	32,34
♠	•			78,9	55,82	65,38	64,87	**36,01**	**46,31**
♠		•	•	81,65	54,72	65,52	81,7	28,77	42,56
♠	•		•	79,15	56,13	65,68	57,97	36,01	44,42
♠	•	•		**81,97**	55,66	**66,30**	**90,3**	23,43	37,2
♠	•	•	•	80,56	55,03	65,39	80,98	23,43	36,34
Átlag				79,91	55,73	65,65	74,19	28,03	39,91

A német részkorpuszon elért eredményeket a 3. táblázatban láthatjuk, ahol a legnagyobb volt a különbség az átlagos nyelvadaptáció eredményei és keresztmérések közt (33,11 F-mérték). Ebben az esetben akkor bizonyult legsikeresebbnek a rendszerünk, amikor a német tanítóhalmazt a 4FX korpusz spanyol részével egészítettük ki, hiszen ekkor 0,59 F-mértékkel jobb eredményt értünk el, mint az indomén eredmény. Az átlagos különbség a szótárillesztés és keresztmérések közt 3,91 volt, de a legjobb esetben a különbség 15,34 volt.

3. táblázat. Kísérleti eredmények a német részkorpuszon. EN: Angol. DE: Német. ES: Spanyol. HU: Magyar. ADAPT: nyelvadaptáció. CROSS: keresztmérések.

Nyelvek				**ADAPT**			**CROSS**		
DE	**EN**	**ES**	**HU**	**Pontosság**	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Szótárillesztés				85,71	7,45	13,71	85,71	7,45	13,71
♠				**64,52**	**41,68**	**50,64**	–	–	–
♠			•	64,4	42,44	51,17	**85,37**	**5,34**	10,06
♠		•		**65,68**	41,98	**51,23**	24,14	13,89	17,64
♠	•			64,48	41,83	50,74	23,26	**25,04**	24,12
♠		•	•	64,48	41,68	50,63	23,36	8,7	12,68
♠	•		•	62,22	41,83	50,03	37,62	23,66	**29,05**
♠	•	•		63,78	42,44	50,97	23,83	10,84	14,9
♠	•	•	•	59,93	**43,36**	50,31	22,93	10,99	14,86
Átlag				63,57	42,22	50,73	34,36	14,07	17,62

A 4. táblázat a spanyol nyelvű eredményeket mutatja. A nyelvadaptációs módszerünk akkor bizonyult a leghatékonyabbnak, amikor a tanítóhalmazt a német részkorpusszal egészítettük ki. Az eredmények azt mutatják, hogy a nyelvadaptáció elsősorban a pontosságra volt hatással, mivel ennek segítségével átlagosan 2,75 ponttal magasabb pontosságot értünk el az indomén mérésekhez képest. Ez a különbség akkor volt a legjelentősebb (4,60), amikor az angol és magyar részkorpuszok uniójával egészítettük ki a tanítóhalmazt. A nyelvadaptáció ebben az esetben is jelentősen meghaladta a szótárillesztés eredményeit, mivel annál 40,28 F-mértékkel jobb eredményt ért el, valamint az átlagos keresztmérések eredményeit is 19,05 F-mértékkel haladta meg.

A magyar részkorpuszon elért eredményeket a 5. táblázat mutatja be. A nyelvadaptáció ugyanazt a 65,25 F-mértéket érte el különböző pontosság és fedés mellett, amikor az angol részkorpusszal, valamint a német és spanyol részkorpuszok uniójával egészítettük ki. A nyelvadaptáció átlagos eredménye és az indomén átlagos eredmények különbsége 0,30 F-mérték, de módszerünk akkor tűnik hatékonyabbnak, ha a nyelvadaptáció során nem csak egy nyelvet használunk. Ugyanakkor a szótárillesztés érte el a legmagasabb pontosságértéket ezen a részkorpuszon 85,86% pontossággal, de a keresztmérések átlagosan 3,13 F-mértékkel magasabbak a szótárillesztésnél.

4. táblázat. Kísérleti eredmények a spanyol részkorpuszon. EN: Angol. DE: Német. ES: Spanyol. HU: Magyar. ADAPT: nyelvadaptáció. CROSS: keresztmérések.

Nyelvek				ADAPT			CROSS		
ES	**EN**	**DE**	**HU**	**Pontosság**	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Szótárillesztés				54,99	31,78	40,28	54,99	31,78	40,28
♠				**62,99**	**45,59**	**52,90**	–	–	–
♠			•	62,01	44,56	51,86	**51,41**	31,39	**38,98**
♠		•		65,32	45,36	**53,54**	32,47	30,13	31,25
♠	•			66,42	44,22	53,09	37,48	27,95	32,02
♠		•	•	65,21	45,02	53,26	34,62	31,84	33,17
♠	•		•	**67,59**	43,87	53,21	42,33	27,84	33,59
♠	•	•		66,77	43,87	52,95	36,32	**32,07**	34,06
♠	•	•	•	66,86	43,64	52,81	37,28	31,73	34,28
Átlag				65,74	44,36	52,96	38,84	30,42	33,91

5. táblázat. Kísérleti eredmények a magyar részkorpuszon. EN: Angol. DE: Német. ES: Spanyol. HU: Magyar. ADAPT: nyelvadaptáció. CROSS: keresztmérések.

Nyelvek				ADAPT			CROSS		
HU	**EN**	**DE**	**ES**	**Pontosság**	**Fedés**	**F-mérték**	**Pontosság**	**Fedés**	**F-mérték**
Szótárillesztés				85,86	22,25	35,34	85,86	22,25	35,34
♠				**80,06**	**54,32**	**64,72**	–	–	–
♠			•	80,21	54,2	64,69	44,74	21,66	29,19
♠		•		79,38	54,44	64,58	45,67	**51,12**	48,24
♠	•			**80,64**	54,79	**65,25**	55,36	44,62	**49,41**
♠		•	•	80,13	**55,03**	**65,25**	46,12	22,49	30,23
♠	•		•	80,27	54,79	65,13	**67,17**	**21,07**	32,07
♠	•	•		80,16	54,91	65,18	49,48	44,73	46,99
♠	•	•	•	80,05	54,79	65,05	39,13	28,76	33,15
Átlag				80,12	54,71	65,02	49,67	33,49	38,47

5. Diszkusszió

Ebben a részben részletesen elemezzük a nyelven belüli és nyelvek közti, illetve nyelvadaptációval elért eredményeinket.

5.1. Nyelven belüli eredmények

Gépi tanuló megközelítésünk a 4FX korpusz minden egyes nyelve esetében jelentősen jobb eredményt ért el a szótárillesztési módszernél, ami igazolja, hogy a szintaxisra épülő gépi tanuló módszer hatékonyan működik az FX-ek automatikus azonosításában különféle nyelveken.

Módszerünk érdekes módon jobban teljesít az angol és magyar nyelveken, mint a német és spanyol nyelveken. Ennek valószínűleg a kevésbé hatékony függőségi elemzés lehet az oka, mivel a jelöltkinyerő módszer a lehetséges FX-eknek csak kisebb hányadát tudta azonosítani a németben és a spanyolban: a német FX-ek 29,01%-ánál és a spanyol FX-ek 25%-ánál a parser nem talált közvetlen szintaktikai kapcsolatot az FX-ige és főnév között. Ezzel szemben ez az arány a magyarban és angolban pusztán 10% körül található.

Általában véve is a német FX-ek azonosítása bizonyult a legnehezebb feladatnak, hiszen a szótárillesztés és a nyelvek közti mérések sem érték el a 20-as F-mértéket. Ezt feltehetőleg annak tudhatjuk be, hogy a németben viszonylag magas a csupán egyszer előforduló FX-ek aránya, ami a szótárillesztés által elért fedési értéket is erősen befolyásolja. Mindezek mellett a német FX-igék voltak a legváltozatosabbak a korpuszban, összesen 93 különböző FX-igét találhatunk a kézzel annotált FX-ekben. Ez a gépi tanulás eredményességére is kihatással volt, hiszen ezekben az esetekben kevés tanító példával találkozott a rendszer egy-egy adott szerkezetre vagy igére nézve.

5.2. Nyelvközi és nyelvadaptációs eredmények

A nyelvek közti mérések eredményei meghaladták a szótárillesztés által elért eredményeket, ami arra világít rá, hogy egy más nyelven tanított gépi tanuló modell hatékonyabbnak bizonyul, mint a célnyelvi szótárillesztés. Ez főleg annak köszönhető, hogy az elérhető szótárak mérete korlátozott volt, így alacsonyabb fedést láthattunk a szótárillesztés esetében, azonban a pontossági értékek kielégítők voltak. Ez alól egy kivételt találunk: a spanyol esetén a szótárillesztés jobban teljesített, mint a nyelvek közti mérések, főként a magas fedési értéknek köszönhetően. Ez egyrészt a nagyobb szótárméretnek tulajdonítható, másrészt pedig a szótárépítési elveknek, hiszen a szótár a lexikai függvények elméleti hátterén alapszik (vö. pl. [23]), és a 4FX korpusz annotációs elvei is részben a lexikai függvényekre hagyatkoznak.

A nyelvek közti eredmények részletesebb vizsgálata rámutat, hogy a magyar korpuszrészen tanított modell elsősorban a pontosságra volt jó hatással, míg a német modell a fedést javította. Ez valószínűleg annak köszönhető, hogy a német korpuszrészben szerepel a legtöbb fajta FX-ige, így a német adatok sokféle példát

tudnak nyújtani a lehetséges FX-ekre, és így a kevésbé gyakori célnyelvi FX-ek megtalálására is részben megoldást ad.

A 2. táblázat alapján elmondhatjuk, hogy a nyelvadaptáció minden esetben felülmúlta az angol nyelven belüli eredményeket. Mivel az angol korpuszrész tartalmazza a legkevesebb FX-et, nem meglepő, hogy a gépi tanuló modell jobb eredményt képes elérni, ha a tanító halmazba több példa kerül, még ha ezek más nyelvből származnak is.

Ha a különféle nyelveken elért eredményeket vetjük össze egymással, látszik, hogy a német alapvetően különbözik a többi nyelvtől. Itt a nyelvközi mérések nyújtották a legalacsonyabb teljesítményt, elsődlegesen a gyenge fedési értékeknek köszönhetően. Ez összefüggésben állhat a korábban már említett okokkal, nevezetesen, hogy a németben nagyon magas az egyszer előforduló FX-ek aránya, továbbá itt a legváltozatosabbak az FX-igék a négy vizsgált nyelv közül. Így tehát a más nyelvű adatokon tanított gépi tanuló modellek nem képesek megfelelő fedést elérni, mivel nincsenek olyan nagyon gyakori FX-ek, melyek lefednék az adatok jelentős hányadát. Az is látszik az adatokból, hogy az angol és magyar korpusz unióján tanított modell teljesít a legjobban a nyelvközi méréseket tekintve a német esetében. Ez a két nyelv jellegzetességeinek köszönhető: amikor csak a magyar adatokon tanítottunk, akkor értük el a legmagasabb pontosságot (85,37%), és a legjobb fedést (25,04%) akkor értük el, amikor csak angol adatokon tanítottunk.

A spanyol eredményeket tekintve észrevehetjük, hogy a legjobb fedési értéket a nyelven belüli mérés eredményezte, így a spanyol FX-ek megtalálása más nyelvű adatok alapján nehéznek bizonyul: csupán a pontossági értékek javulnak a más nyelvű tanító adatok használatával. Valószínűleg ebben a tekintetben a spanyol a némethez hasonlít: a spanyolban is viszonylag magas az egyszer előforduló FX-ek és FX-igék aránya, így a más nyelvű adatok nem tudták segíteni a gépi tanuló eljárást a ritka példák megtalálásában. Továbbá, a nyelvadaptáció eredményei is átlagosan csak 2,75 százalékponttal magasabbak, mint a nyelven belüli mérés esetében.

Ami a magyart illeti, a legsikeresebb nyelvközi kísérletnek az angol mint forrásnyelv alkalmazása bizonyult, míg az angol és spanyol adatok uniója adta a legmagasabb pontosságot. Ezt az magyarázza, hogy az angol modell is nagyon magas pontosságot ért el a nyelven belüli kísérlet során is, így az angol adatokból a modell meg tudja tanulni, hogyan válassza ki a jelöltekből a tényleges FX-eket. Mindemellett, a német adatokon tanított modell magas fedési értéket képes elérni, valószínűleg a korpuszban levő FX-ek változatossága miatt.

6. Összegzés

Ebben a munkában bemutattuk nyelvfüggetlen eljárásunkat félig kompozicionális szerkezetek azonosítására. Módszerünk első lépésben a lehetséges jelölteket nyeri ki a szövegekből szintaktikai jellemzőkre építve, majd egy gépi tanuló modell kiválasztja ezek közül a tényleges FX-eket. Eljárásunkat a 4FX korpuszon teszteltük.

A legtöbb esetben a gépi tanuláson alapuló keresztmérésekkel néhány százalékponttal jobb eredményt sikerült elérni, mint a célnyelvi szótárillesztés segítségével, például az angol nyelv esetében a különbség 8 százalékpontnyi az F-mértéket tekintve. Ez azt mutatja, hogy a gépi tanuló megközelítésünk még akkor is hatékonyabb az egyszerű szótárillesztésnél, ha a tanító halmaz és a teszthalmaz eltérő nyelvű. A nyelvadaptációval elért eredmények megközelítik, sőt bizonyos esetekben meg is haladják 0,5-1 százalékponttal a tízszeres keresztvalidációval elért eredményeket: például az angol nyelv esetében a legjobb eredményt a spanyol–német adathalmazról adaptálva értük el. Mindez arra utal, hogy a nyelvadaptációs technikák sikeresen alkalmazhatók a többszavas kifejezések automatikus azonosításában, különösen akkor, ha a célnyelven csak kis mennyiségű annotált adat áll rendelkezésre.

A későbbiekben szeretnénk az egyes nyelvek sajátságaira építve jellemzőinket bővíteni és a módszert más nyelvekre is kiterjeszteni.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Vincze, V., Nagy T., I., Farkas, R.: Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In: Proceedings of ACL 2013, Sofia, Bulgaria, ACL (2013) 255–261
2. Rácz, A., Nagy T., I., Vincze, V.: 4FX: Light Verb Constructions in a Multilingual Parallel Corpus. In: Proceedings of LREC'14, Reykjavik, Iceland, ELRA (2014)
3. Tan, Y.F., Kan, M.Y., Cui, H.: Extending corpus-based identification of light verb constructions using a supervised learning framework. In: Proceedings of MWE 2006, Trento, Italy, ACL (2006) 49–56
4. Stevenson, S., Fazly, A., North, R.: Statistical Measures of the Semi-Productivity of Light Verb Constructions. In: MWE 2004, Barcelona, Spain, ACL (2004) 1–8
5. Van de Cruys, T., Moirón, B.n.V.: Semantics-based multiword expression extraction. In: Proceedings of MWE 2007, Morristown, NJ, USA, ACL (2007) 25–32
6. Vincze, V.: Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses. PhD thesis, University of Szeged, Szeged, Hungary (2011)
7. Diab, M., Bhutada, P.: Verb Noun Construction MWE Token Classification. In: Proceedings of MWE 2009, Singapore, ACL (2009) 17–22
8. Nagy T., I., Vincze, V., Berend, G.: Domain-Dependent Identification of Multiword Expressions. In: Proceedings of RANLP 2011, Hissar, Bulgaria, RANLP 2011 Organising Committee (2011) 622–627
9. Vincze, V., Nagy T., I., Zsibrita, J.: Félig kompozicionális szerkezetek automatikus azonosítása magyar és angol nyelven. In Tanács, A., Vincze, V., eds.: VIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2011) 59–70

10. Tu, Y., Roth, D.: Learning English Light Verb Constructions: Contextual or Statistical. In: Proceedings of MWE 2011, Portland, Oregon, USA, ACL (2011) 31–39
11. Nagy T., I., Vincze, V., Farkas, R.: Full-coverage Identification of English Light Verb Constructions. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, Asian Federation of Natural Language Processing (2013) 329–337
12. Bohnet, B.: Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of Coling 2010. (2010) 89–97
13. Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., Nivre, J.: The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In: Proceedings of the Twelfth Conference on Computational Natural Language Learning, Association for Computational Linguistics (2008) 159–177
14. Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G., Uszkoreit, H.: TIGER: Linguistic interpretation of a German corpus. Research on Language and Computation **2**(4) (2004) 597–620
15. Marimon, M., Fisas, B., Bel, N., Villegas, M., Vivaldi, J., Torner, S., Lorente, M., Vázquez, S., Villegas, M.: The IULA Treebank. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), Istanbul, Turkey, European Language Resources Association (ELRA) (2012) 1920–1926
16. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. (2013) 763–771
17. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of LREC 2010. (2010)
18. McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., Lee, J.: Universal dependency annotation for multilingual parsing. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria, Association for Computational Linguistics (2013) 92–97
19. Vincze, V.: Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In: Proceedings of LREC 2012, Istanbul, Turkey (2012)
20. Krenn, B.: Description of Evaluation Resource – German PP-verb data. In: Proceedings of MWE 2008, Marrakech, Morocco (2008) 7–10
21. Kolesnikova, O., Gelbukh, A.: Supervised machine learning for predicting the meaning of verb-noun combinations in Spanish. In: Advances in Soft Computing. Springer (2010) 196–207
22. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**(1) (2009) 10–18
23. Mel'čuk, I.: Esquisse d'un modèle linguistique du type "Sens<->Texte". In: Problèmes actuels en psycholinguistique. Colloques inter. du CNRS, no. 206, Paris, CNRS (1974) 291–317